Theseuksen kehittäminen
Katsaus historiaan, kurkistus tulevaan

Theseus 10-vuotisjuhlaseminaari 21.11.2018

Samu Viita (samu.viita@helsinki.fi)

Esityksen runko

teknispainotteinen katsaus Theseuksen historiaan ja tulevaisuuteen

- Theseuksen suunnittelu, pilotointi ja tuotannon ensivaiheet
- Pikakelaus nykypäivään: poimintoja matkan varrelta
- Kansalliskirjaston julkaisuarkistopalvelun kehittämisperiaatteita
- Tuoreita ominaisuuksia
- Kurkistus Theseuksen tulevaisuuteen

Theseuksen suunnittelu - lähtötilanne

- 11. joulukuuta 2007 rehtorit allekirjoittivat sopimuksen ammattikorkeakoulujen yhteisestä verkkokirjastopalvelusta
- Kansalliskirjasto (KK) valittiin verkkokirjaston toteuttajaksi
  - KK julkaisuarkisto Doriaan kuuluneet E-thesis, LutPub ja Stadia toimivat referenssinä opinnäytteiden julkaisuprosessista.

Theseuksen suunnittelu ja pilotointi

- Palvelusopimus AMK:ien ja KK:n välillä vuosiksi 2008 - 2013
- Suunnittelu- ja kehitystyö tapahtui talven ja kevään 2008 aikana.
- Pilotointi touko - joulukuussa 2008
  - 11 AMK:ia mukana pilotissa (Arcada, HAAGA-HELIA, Jyväskylä, Kemi-Tornio, Lahti, Laurea, Metropolia, Pirkanmaa, Satakunta, Seinäjoki ja Turku).
- Pilotoivien amkien kanssa kokoonnuttiin suunnitteluvaiheessa useaan otteeseen ja hahmoteltiin vaatimuksia.

Theseuksen suunnittelu

- Avoimen lähdekoodin sovellus DSpace valittiin myös Theseukseen
- Suunnittelupalavereissa nousi erityisesti esiin **itsetallentamiseen liittyvän käytettävyyden** tärkeys.
- Palvelun **juridiset seikat** askarruttivat
  - Julkaisulupa, uudelleenkäyttöön liittyvät lisenssit, käyttöehdot, metadata / kokoteksti...
- AMK:it tilasivat lakitoimistolta apua. -> Theseuksen ominaisuuksiin merkittävä lisäys viime hetkellä:
- Creative Commons lisensointi opinnäytteisiin.

Kehittämisvaiheen haasteita

- Miten toteuttaa kaikki vaatimukset neljässä kuukaudessa tuotantokuntoon puolen htv:n resursseilla:
- **Syöttölomakkeen** vaatimukset opinnäytteiden luovuttamiseen:
  - Kriittinen osa valmistumisprosessia, ei siedä katkoja tai hidasteluja
  - Ennustettu vuosikartunta jopa 30 000 opinnäytettä
  - Käyttäjäystävällisyys välttämätöntä -> Lomaketta käytetään keskimäärin kerran elämässä
  - Haka-kirjautuminen + tietojen esitäyttö
  - Tuki eri selaimille
  - Monikielisyystuki

Kehittämisvaiheen haasteita

- Syötön vaatimukset eivät toteutuneet Dspace-lomakkeella 2008:
  - Ei kielikäännöksiä, metadatan kieliversiointia, kenttien esitäyttöä, URN-tunnuksen automaattista hakua yms…
- **Päätös**: Opinnäytteiden luovutuslomake itsenäisenä sovelluksena ja eri palvelimelle kuin julkaisuarkisto
  - Syöttölomake räätälöitävissä paremmin ja luovutus riippumaton Theseuksen julkaisuarkiston mahdollisista katkoista

Kehittämisvaiheen haasteita

- **Dspacen muokkaus verkkoarkistolle (Theseukselle) sopivaksi**
  - Dspace suunniteltu alun perin yhden organisaation tarpeisiin, vaati muokkauksia koodiin
    - Ulkoasu, ohjeistus, kielikäännökset
    - Joitain Dspacen ominaisuuksia tuli piilottaa käyttäjiltä
    - Puutteiden ja bugien, kuten rajapintojen tietoturvavuotojen korjaaminen
    - Dspacen käyttäjänhallinnan sovittaminen Theseuksen käyttöön

Kehittämisvaiheen haasteita

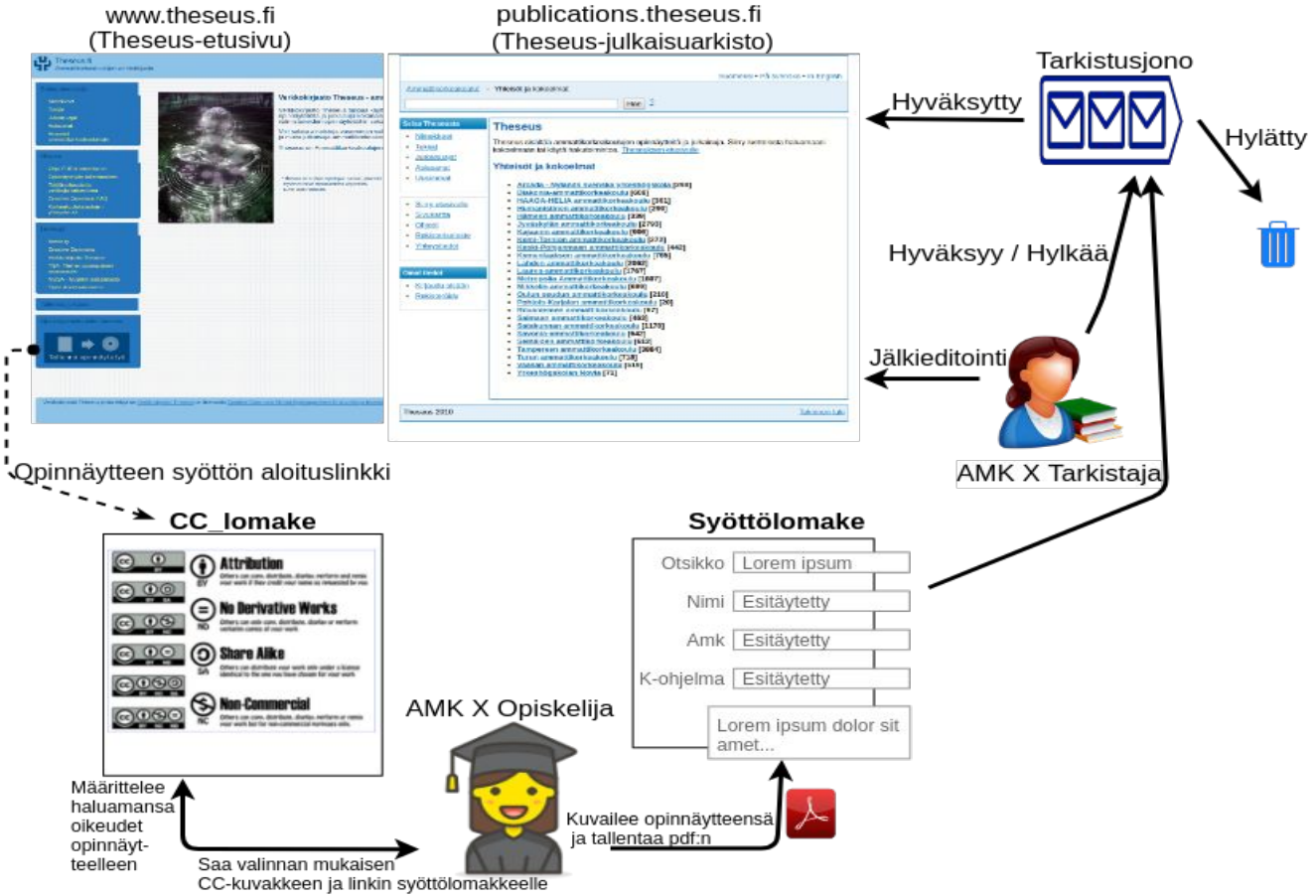
- **Haka-käyttäjätunnistusjärjestelmä**
  - Hakan ideana tarjota kertakirjautuminen verkoston jäsenille
  - Teoriassa Hakan luottamusverkostosta tulevat tiedot ovat yhteismitallisia
    - Selvitys paljasti kuitenkin toista: AMK-kohtaisia vaihteluja käytänteissä oli attribuuttien ja niiden arvojen osalta -> Vaati yhtenäistämistä AMK:ien IdP-vastaavien kanssa
  - Osaa atribuuttien käytön poikkeuksista ei saatu yhteismitalliseksi
    - Koodissa AMK-kohtaisia poikkeuksia niiden hallintaan
  - Osa AMK:eista oli pitkään ilman Haka-tekniikkaa
    - Tuli toteuttaa lomake myös ilman Hakaa kirjautuville

Kehittämisvaiheen haasteita

- Vaatimusmäärittelyyn viime hetken lisäys CC-lisenssien tukemisesta
  - Syötön yhteyteen haluttiin tuki käyttäjän valintoja ohjaavalle lisenssivalitsimelle, joka tuottaisi metadataan **ja** pdf:ään CC-lisenssin kuvineen ja linkkeineen
  - Pdf-tiedostoon lisäyksestä kuitenkin luovuttiin - se koettiin liian riskialtiiksi
    - syöttövolyymi suuri, pdf:ien versiovariaatiota odotettavissa paljon
  - Ratkaisu: Syöttö kaksivaiheinen, ensin lisenssin valinta, toisessa vaiheessa itse opinnäytteen kuvailu
    - Mahdollistaa lisenssitiedon lisäämisen opinnäytteeseen näiden välissä

Täsmentynyt aikataulu

- Theseus Shibboleth autentikoidulla syötöllä tuotantokunnossa 16.5.2008
- Creative Commons tuki kesän 2008 aikana
- Aikataulu oli siis edelleen haastava.
  - Hätäapua koodaukseen KK:n Nelli-tiimiltä
- Lopulta Theseus saatiin onnellisesti tuotantokuntoon aikataulun mukaisesti!


www.theseus.fi
(Theseus-etusivu)
publications.theseus.fi
(Theseus-julkaisuarkisto)
Tarkistusjono
Hyväksytty
Hylätty
Hyväksyy / Hylkää
Jälkieditointi
AMK X Tarkistaja
Qpinnäytteen syöttön aloituslinkki
CC_lomake
Attribution
No Derivative Works
Share Alike
Non-Commercial
Syöttölomake
Otsikko
Lorem ipsum
Nimi
Esitäytetty
Amk
Esitäytetty
K-ohjelma
Esitäytetty
Lorem ipsum dolor sit amet...
AMK X Opiskelija
Määrittelee haluamansa oikeudet opinnäyt-teelleen
Saa valinnan mukaisen CC-kuvakkeen ja linkin syöttölomakkeelle
Kuvailee opinnäytteensä ja tallentaa pdf:n

Vastuunjaot / palvelimet
2008 - 2014

Kansalliskirjasto
Yritys

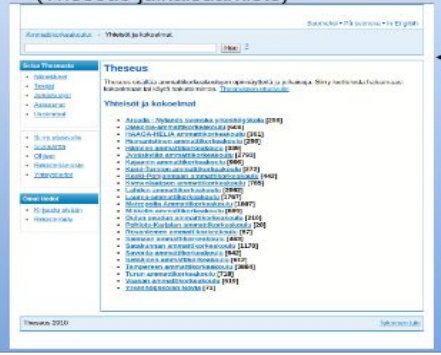
WWW-palvelin
(Verkkoasema, Proactum)

www.theseus.fi
(Theseus etusivu)

Dspace-palvelin
(Kansalliskirjasto)

publications.theseus.fi
(Theseus-julkaisuarkisto)

Tarkistusjono



Syöttölomakepalvelin
(Kansalliskirjasto)

CC-lomake

Syöttölomake

Otsikko Lorem ipsum
Nimi Esitäytetty
Amk Esitäytetty
K-ohjelma Esitäytetty
Lorem ipsum dolor sit amet...

Tuotannon ensivaiheet

- Day one: 0 tietuetta
- Opinnäytteiden määrä nousi kesän aikana yli 1000 kpl
  - Takautuvia aineistoja vietiin Theseukseen SAMK:in toimesta jo kesän aikana
- Asiakastuki aluksi haastavaa: paljon suoria yhteydenottoja asiakkailta
  - AMK-kirjastot, opiskelijat, tiedonhakijat, opettajat…
  - Tilanne helpottui Kansalliskirjastossa, kun **Theseus-rukkaset** aloittivat toimintansa 2010: Suorat asiakaskontaktit ohjattiin rukkasille

Tuotannon ensivaiheet

- Syöttöprosessi osoittautui toimivaksi
  - Kesti kuormaa hyvin, myös pilotin jälkeen
  - Syöttölomake nopea ja käytännössä vapaa katkoista tai hidasteluista
  - Syöttö vastasi odotuksia myös käytettävyyden osalta
- Pilotin aikana käyttöluvut ja kartunta maltillista, ei palvelun kuormaan liittyviä haasteita vielä 2008

Ylläpidon haasteita 2009 ->

- Käyttökuorma omaa luokkaansa
  - Opinnäytteitä hyväksyviä virkalijoita lähes 200 (arvio)
  - Opinnäytteitä syötetään ruuhka-aikaan parhaillaan yli 300 päivässä
  - Aktiivisia käyttäjiä sivustolla yli 2000 (20.11.2018 Google analytics)
  - Paljon hakurobottiliikennettä
  - Palvelunestohyökkäykset

Pikakelaus nykypäivään

- **2009** Julkaisujen syöttömahdollisuus: Julkaisutyöryhmä suunnitteli metadatan. KK-kehitti pdf-latauksien tilastointiohjelma Simplestatsin julkaisuarkistoihinsa
- **2010** AMK:it lähes täysilukuisina mukana, vuosikartunta n. **10 000**, pdf-latauksia n. **1,5 milj.**
- **2011**: Toinen täysipäiväinen työntekijä julkaisuarkistojen tekniikkaan KK:ssa (**Päivi Rosenström**)! KK otti virtuaalipalvelintekniikan ja versionhallinnan käyttöön.
- **2012:** Theseuksen “etusivu” www.theseus.fi sai uuden osoitteen submission.theseus.fi ja www.theseus.fi vapautui julkaisuarkiston käyttöön.
- **2013**: Vuosikartunta n. **15 000,** pdf-latauksia n. **12 milj.**
- **2014:** Theseuksen “etusivu” eli nykyinen submission.theseus.fi siirtyi KK:n ylläpitoon.
- **2016**: Kolmas täysipäiväinen työntekijä julkaisuarkistojen tekniikkaan KK:ssa (**Anis Moubarik**)! **2017**: Responsiivinen käyttöliittymä, kansikuvat ja Rest-rajapinta. Vuosikartunta n. **17 500,** pdf-latauksia n. **16,5 milj**.

Kansalliskirjaston julkaisuarkistojen kehittämisperiaatteita

- KK:n julkaisuarkistopalvelulla nykyisin 11 julkaisuarkistoa ylläpidossa
  - Theseus, Doria, Julkari, Jukuri, Valto, Lauda, TamPub, UTUPub, LUTPub, Osuva, Fenno-Ugrica
- Asiakaskohtaisia kehitystoiveita riittää - Suurin osa kuitenkin yhteisiä
- Ominaisuustoiveet pyritään toteuttamaan kaikille aina kun mahdollista
  - Kaikki saavat hyödyn uudistuksista ja bugikorjauksista

Tärkeimmät kehittämistä ohjaavat periaatteet KK:n julkaisuarkistopalveluissa

- Tietoturva
- Vakaus
- Saatavuus ja saavutettavuus: Mahdollisimman monen käytettävissä
  - Käyttö erilaisilla päätelaitteilla ja apuvälineillä (esim. näkövammaiset)
  - Myös koneet käyttäjinä
- Tiedon uudelleenkäytön mahdollistaminen
  - Julkaisuarkistot osana isompaa IT-infrastruktuuria (Jyrki Ilvan esitys kertoo tästä tarkemmin)

Tärkeimmät kehittämistä ohjaavat periaatteet

- Edellä luetellut periaatteet tarkoittavat usein myös maltillisuutta uusien toimintojen tai esim. grafiikan suhteen

"keep it simple"

Theseuksen tuoreita ominaisuuksia 2017 - 2018

- Responsiivinen käyttöliittymä
  - Mobiililaitteille tuki
- Kansikuvat
  - Dspace 5:en myötä ImageMagick-ohjelma hoitaa
- Uusi REST-rajapinta
  - Dspacen virallinen, korvasi aiemmin KK:ssa laaditun REST:in
- Refworks-exportointi
- äänen ja videon streamaus HTML5 -tekniikalla
- Kuvagalleriatoiminto

Kurkistus Theseuksen lähitulevaisuuteen

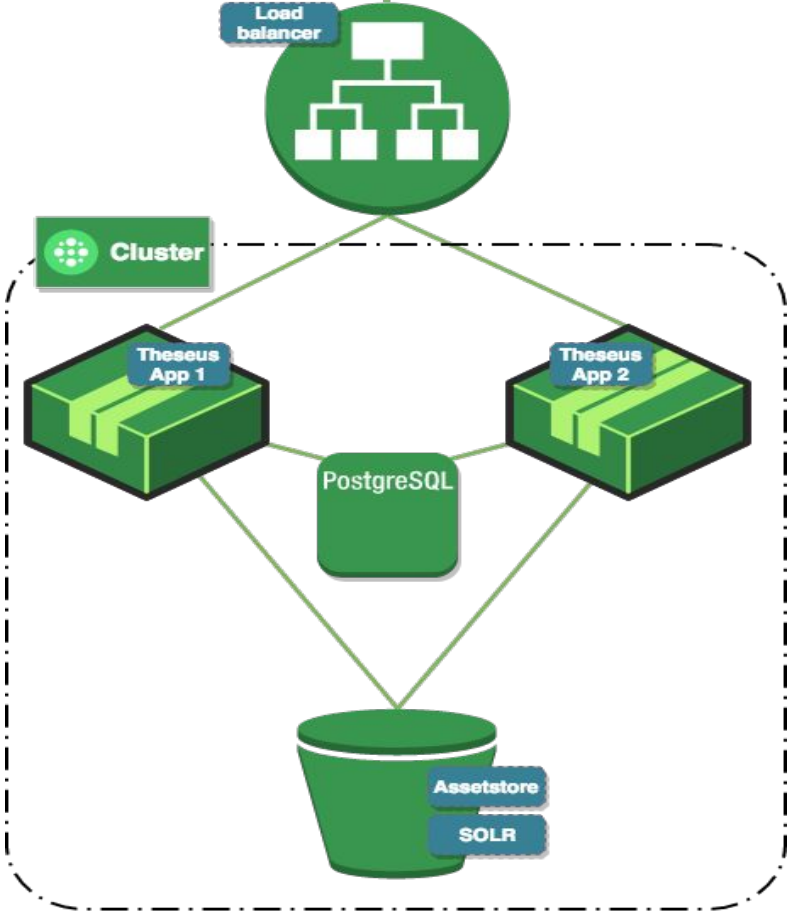
- Luovutaan vanhasta Theseuksen syöttölomakkeesta
  - tilalle KK:n muokkaama Dspace-lomake
  - Miinuksia toistaiseksi:
    - Ei metadataan kielimääreitä eikä tietojen esitäyttöä Hakan perusteella
  - Hyötyjä:
    - Finto-kytkös, jatka myöhemmin -toiminto, toistuvien kenttien syöttö helpompaa
    - Lomakkeiden kustomointi ja päivittäminen helppoa
    - Uusia ominaisuuksia Theseuksen syöttölomakkeelle tulevaisuudessa
    - Kirjaston henkilökunnalla mahdollisuus siirtyä kokonaan Haka-kirjautumiseen (vaatii Haka-attribuuttien osalta lisäselvitystä / koordinointia)

Kurkistus lähitulevaisuuteen

- Kansallisen metadatasuosituksen käyttöönotto (Tanja Vienonen, KK)
- Kokoelmien karsinta: Yksi opinnäyte- ja julkaisukokoelma per AMK
  - Koulutusohjelma- yms. tiedot esiin selausvaihtoehtojen ja fasettien avulla
- Ulkoasu-uudistus (Minna Marjamaan ja Tiina Tolosen esitys)
- EU:n vaatimukset: Saavutettavuuden ja GDBR:n huomioonotto entistä paremmin
- Endnote-eksportointi
  - Odottaa metadatauudistusta ja viittaussuosituksen valmistumista
- APA-tyylinen viittauslaatikko
  - Odottaa metadatauudistusta ja viittaussuosituksen valmistumista

Kurkistus lähitulevaisuuteen

- Kuorman tasaus:
  - Kuormaa tasattu aiemmin vain erottamalla syöttö ja julkaisuarkisto (Dspace) eri palvelimille
  - Suunnitelmissa hoitaa syöttökin jatkossa Dspacella
  - Tarve kuorman tasaukselle
- DSpace (www.theseus.fi) klusteroidaan:
  - yhteensä viisi palvelinta pyörittämään Theseuksen julkaisuarkistopuolta


Load balancer
Cluster
Theseus App 1
Theseus App 2
PostgreSQL
Assetstore
SOLR

Kurkistus lähitulevaisuuteen

- http -> https
  - Luovutaan http:stä, vaatii vielä selvitystä haravoijien osalta
- **Justus-kytkentä Theseukseen**
  - Julkaisujen syöttöön ja tiedonkeruuseen helpotusta Justuksen avulla
  - Tästä tarkemmin Joonaksen Nikkasen esityksessä!

Alustavia suunnitelmia pidemmälle tulevaisuuteen

- Julkaisuarkistojen tietojen avaaminen linkitettynä datana
  - Metadatauudistus edistää linkitetyn datan käyttöönottomahdollisuuksia
  - Datat todennäköisesti Turtle-formaattia
- Pimeä arkisto Theseuksen yhteyteen
  - Kokoelmarakenteen purku ja Dspace-syöttölomakkeen käyttö avaa mahdollisuuksia pimeän arkiston toteuttamiselle
  - Vaatii vielä teknistä ja muuta selvitystyötä lisää!

Kiitos!

Lopuksi vinkki juhlaseminaarin väelle:
Sulzer-höyrykone
2. kerroksessa!